Eliciting Human Judgment for Prediction Algorithms

Rouba Ibrahim
School of Management, University College London, rouba.ibrahim@ucl.ac.uk

Song-Hee Kim
Marshall School of Business, University of Southern California, songheek@marshall.usc.edu

Jordan Tong
Wisconsin School of Business, University of Wisconsin-Madison, jordan.tong@wisc.edu

August 19, 2020

Even when human point forecasts are less accurate than data-based algorithm predictions, they can still help boost performance by being used as algorithm inputs. Assuming one uses human judgment indirectly in this manner, we propose changing the elicitation question from the traditional direct forecast (DF) to what we call the private information adjustment (PIA): how much the human thinks the algorithm should adjust its forecast to account for information the human has that is unused by the algorithm. Using stylized models with and without random error, we theoretically prove that human random error makes eliciting the PIA lead to more accurate predictions than eliciting the DF. However, this DF-PIA gap does not exist for perfectly consistent forecasters. The DF-PIA gap is increasing in the random error that people make while incorporating public information (data that the algorithm uses) but is decreasing in the random error that people make while incorporating private information (data that only the human can use). In controlled experiments with students and Amazon Mechanical Turk workers, we find support for these hypotheses.

Key words: laboratory experiments, behavioral operations, random error, elicitation, forecasting, prediction, discretion, expert input, private information, judgment, aggregation

1. Introduction

Because of increased access to data and advancements in machine-learning algorithms, a common operational improvement initiative is to replace human forecasters with data-driven prediction algorithms. For example, in our motivating setting, a hospital needs surgery duration forecasts to schedule operating room use, which costs $\$2,190$ per hour on average (Childers and Maggard-Gibbons 2018). Using surgery duration data from that hospital, Ibrahim and Kim (2019) show that physicians' mean absolute percent forecast error was 33%, whereas algorithms based on available patient and surgery data reduced that error to 29%.

Nevertheless, even if humans are not better than algorithms head-to-head, their judgments can still help. In the above example, the hospital could improve predictive accuracy even further, to under 27%, by using the physicians' forecasts as an input (along with the other data) to the algorithm. In other words, the best forecasts often come not from replacing humans with algorithms, but from combining them.

In this research, we ask the following question: If we know that we are going to use human judgment not

directly, but rather *indirectly*, in an algorithm, should we elicit something else besides point forecasts? If so, what human judgment should we elicit, and why might it work better?

We theorize that the primary reason why humans add value to algorithms is that they have access to private information that the algorithm does not use, for example, because it is not in the database or there are not sufficient historical training data for the algorithm to effectively use it. Therefore, we consider whether, rather than asking for a human's *direct forecast* (DF), it may be better to instead ask about her judgment of this private information's impact (even if the system designer does not know what this private information is ahead of time). Specifically, we propose the idea of eliciting the *private information adjustment* (PIA)—how much the human thinks the algorithm should adjust its forecast to account for the information that only the human has.

Using stylized models (§2), we theorize that the PIA leads to more accurate predictions than DF only if there is human random error. That is, from a predictive accuracy perspective, there is no difference between eliciting a DF or the PIA if people are perfectly consistent in how they use information to make a forecast. However, if they are inconsistent, then the PIA should help algorithms more than the DF. Furthermore, the models shed light on how random error creates this difference by predicting which environmental conditions would lead to greater differences in performance. Namely, they show that the PIA's advantage, relative to DF, is larger when "public" data—the data that the algorithm uses—are complicated for the human to process but smaller when the human's private information is complicated to process instead.

To test these hypotheses regarding the difference in performance between DF and PIA, we conducted controlled experiments in which we elicited human judgments for 50 simulated surgery durations based on predictive data. We told our participants that the hospital's algorithm had access to only some of the data ("public information") and that only the participant had access to the other data ("private information"). In one condition, we elicited judgment by asking for the participant's DF for each surgery, while in the other, we elicited the participant's PIA for each surgery. Then, for each condition, we calibrated prediction algorithms using the first 35 surgeries and tested their predictive performance using the last 15 surgeries. In Experiment 1 (§3), conducted with university students and replicated with Amazon Mechanical Turk (MTurk) workers, we find that prediction algorithms performed significantly better when they had access to the participants' PIA as inputs as opposed to their DF: their average root mean squared error (RMSE) in the test sets was 21% lower. In Experiment 2 (§4), we manipulate random error magnitudes by making the public or private information more or less complex: subjects must aggregate multiple factors when the information is complex but are provided one equivalent factor when the information is not complex. Consistent with our theoretical development, we find that the RMSE for PIA is 48% lower than for DF when the public information is complex but only 6% lower than for DF when the private information is

complex. Finally, in §5, we discuss why private information exists in practice, implications of our findings, and opportunities for future research.

We contribute to four main bodies of research. Management science researchers have recognized the potential value in *integrating human judgment with forecasting algorithms* (see Arvan et al. 2019 for a review). Humans often possess so-called "domain knowledge": better and more up-to-date information than what statistical models use (see §3 of Lawrence et al. 2006). Such domain knowledge is the generally accepted explanation for why human judgmental forecasting sometimes even outperforms statistical models in practice (see Lawrence et al. 2000 for sales forecasting and Alexander Jr 1995 for earnings forecasting). The two most common integration approaches are to make judgmental adjustments to an algorithm's point forecast (e.g., see Carbone et al. 1983, Fildes et al. 2009) or to combine separate human and algorithm point forecasts (e.g., see Blattberg and Hoch 1990, Goodwin 2000). We contribute by examining a different human elicitation question from the point forecast. Notably, our proposed method is *not* equivalent to judgmental adjustments because we use the PIA as an *input* for the prediction algorithm. In fact, in our experiments, using PIA responses to adjust algorithm forecast outputs yields poor predictive performance.

A stream of behavioral operations management research studies the *system design implications of human random error*. For example, the best way to design contracts (Su 2008, Ho and Zhang 2008), queues (Huang et al. 2013, Tong and Feiler 2017), or auctions (Davis et al. 2014) changes once the system designer considers human random error. Most closely related to our paper is Kremer et al. (2016). They show that human random error causes eliciting human forecasts in a top-down fashion to be more effective in some environments but bottom-up forecasting to be more effective in others. We contribute by showing how a forecasting system's elicitation design impacts performance once one considers human random error, even if it has no effect without human random error.

Researchers in judgment and decision making have made advancements in developing *strategies to improve human judgment accuracy*. Perhaps the most well-known idea is to harness the "wisdom of crowds" (e.g., see Surowiecki 2005) through averaging multiple people's judgments. Interestingly, because people are so inconsistent (Kahneman et al. 2016), even averaging multiple judgments by the same person separated by time (Vul and Pashler 2008) or with a prompt to think differently (Herzog and Hertwig 2009) helps, albeit only about half as much as averaging judgments by two different people (Mannes et al. 2012). Most closely related to our work in this stream is Palley and Soll (2019), who develop a new elicitation method that improves the "wisdom of crowds" strategy by estimating the amount of shared information between individuals. Our elicitation strategy also seeks to improve an aggregation strategy by addressing the issue of disentangling the shared information between the human and the algorithm.

Lastly, studying the benefit of *incorporating discretion from humans with local knowledge in operational decision making has been an emerging topic of study in operations management. Various application domains have been considered, such as capacity decisions in service operations (Campbell and Frei 2011), sales forecasting (Osadchiy et al. 2013), price setting (Phillips et al. 2015), hospital unit admission decisions (Kim et al. 2015), and product removal decisions in retail stores (Kesavan and Kushwaha 2020).*

2. Theory Development

In this section, we leverage simple mathematical models to compare the theoretical performance of a prediction algorithm that uses human DF and one that uses human PIA. Specifically, our focus is on showing that the difference in performance depends critically on whether or not we assume the forecaster suffers from random error. Our theoretical development is agnostic about the exact sources of this random error and does not attempt to provide a detailed description of the psychology of prediction. Rather, the point of the models is to clearly demonstrate that including random error in the forecast is sufficient to generate differences between DF and PIA. We use these results to motivate two hypotheses about whether and how eliciting the PIA will be more effective than DF.

2.1. Surgery Duration Assumptions

We assume an actual surgery duration, Y, is a random variable defined by the linear model

$$Y = v + \sum_{i \in \mathcal{P} \cup \mathcal{I}} w_i X_i + \epsilon, \tag{1}$$

where we separate the public factors, denoted by the index set $\mathcal{P}$, from the private factors, denoted by the index set $\mathcal{I}$. In (1), ϵ is an error term, with $\mathbb{E}[\epsilon] = 0$, which represents true environmental random shocks, i.e., random variations that are impossible to predict even with all public and private information. We assume that ϵ and $(X_i)_{i \in \mathcal{P} \cup \mathcal{I}}$ are mutually independent.

2.2. DF and PIA Are Equivalently Effective with Consistent Forecasters (No Random Error)

We define DF and PIA for the *consistent forecaster* who does not suffer from random error as follows:

$$DF^* = v^* + \sum_{i \in \mathcal{P} \cup \mathcal{I}} w_i^* X_i \text{ and } PIA^* = \sum_{i \in \mathcal{I}} w_i^* X_i. \tag{2}$$

Here, we assume that v^* and w_i^*, for $i \in \mathcal{P} \cup \mathcal{I}$, are *deterministic*, though not necessarily known by the algorithm a priori. (Note that they can be any constants and are not necessarily "optimal". For example, setting $w_i^* = 0$ is equivalent to assuming humans do not use that information.) Define the best-fitting models of Y given the public factors and DF^* or PIA^* using linear regression:

$$\text{(Model-}DF^*\text{)} \quad M_{DF^*} = \alpha_0 + \sum_{i \in \mathcal{P}} \alpha_i X_i + \beta_{DF^*} DF^*, \tag{3}$$

$$\text{(Model-}PIA^*\text{)} \quad M_{PIA^*} = \gamma_0 + \sum_{i \in \mathcal{P}} \gamma_i X_i + \beta_{PIA^*} PIA^*. \tag{4}$$

Then, the following proposition holds. We relegate the proofs of all propositions to the Appendix.

PROPOSITION 1. *Model-DF^* and Model-PIA^* yield the same predictions.*

That is, with consistent forecasters, predicting surgery durations using DFs as model inputs yields the same predictions as using PIAs as model inputs; the two elicitation methods are equivalent from the algorithm's perspective.

2.3. PIA Outperforms DF with Inconsistent Forecasters (Random Error)

Next, we define DF and PIA for the *inconsistent forecaster* who does suffer from random error:

$$DF^b = v^b + \sum_{i \in \mathcal{P} \cup \mathcal{I}} W_i^b X_i \text{ and } PIA^b = \sum_{i \in \mathcal{I}} W_i^b X_i. \tag{5}$$

Here, we assume that W_i^b are *random variables* with $\mathbb{E}[W_i^b] = \bar{w}_i^b$ and $\text{Var}[W_i^b] > 0$. We also assume that W_i^b and X_i are all mutually independent, for $i \in \mathcal{P} \cup \mathcal{I}$. Thus, in contrast to (2), (5) captures inconsistencies or random error in assigning weights to each factor. For example, these random weights could reflect inconsistencies in the encoding of information, memory retrieval, aggregation of multiple factors, or the translation from one domain to another. Also, note that this random weights model can capture ideas such as inconsistency in which factors humans take into account or adding a random term to DF^b but not to PIA^b. For the purposes of this paper, we make no strong claim about the psychological source of this random error—only that it exists (e.g., see Kahneman et al. 2016) and is greater when people are asked to account for more factors.

Define the best-fitting models of Y given the public factors and DF^b or PIA^b using linear regression:

$$\text{(Model-}DF^b\text{)} \quad M_{DF^b} = \alpha_0 + \sum_{i \in \mathcal{P}} \alpha_i X_i + \beta_{DF^b} DF^b, \tag{6}$$

$$\text{(Model-}PIA^b\text{)} \quad M_{PIA^b} = \gamma_0 + \sum_{i \in \mathcal{P}} \gamma_i X_i + \beta_{PIA^b} PIA^b. \tag{7}$$

In contrast to the equivalence result in Proposition 1 for the consistent-forecaster model, the following proposition demonstrates the benefit of eliciting the PIA compared to eliciting the DF under the inconsistent-forecaster model. (Note that all propositions hold for both MSE and RMSE; we use RMSE to report our experiment results.)

PROPOSITION 2. *The mean squared error (MSE) for predictions under Model-DF^b is strictly larger than that under Model-PIA^b, i.e., $\mathbb{E}[(Y - M_{DF^b})^2] > \mathbb{E}[(Y - M_{PIA^b})^2]$.*

The intuition is that from the algorithm's perspective, the value of the human input is the private information—the algorithm already has the public information. The algorithm can infer the private information equally well from DF or PIA responses without human random error. However, when there is human random error, the algorithm can more accurately infer the private information from the PIA. Based on this result, we formulate our first hypothesis.

Hypothesis 1 *All else equal, a prediction model that is calibrated using DF yields less accurate predictions than a prediction model that is calibrated using PIA.*

2.4. The DF-PIA Gap Magnitude Depends on Random Error Location

Proposition 2 establishes our main result that because of human random errors, using PIA yields more accurate predictions than using DF. We now investigate how the "location" of these random errors (i.e., whether they occur incorporating public versus private factors) affects the performance difference between DF and PIA. To do so, we study the behavior of the *DF-PIA gap*, which we define as the difference in the MSEs from Proposition 2, $\mathbb{E}[(Y - M_{DF^b})^2] - \mathbb{E}[(Y - M_{PIA^b})^2]$.

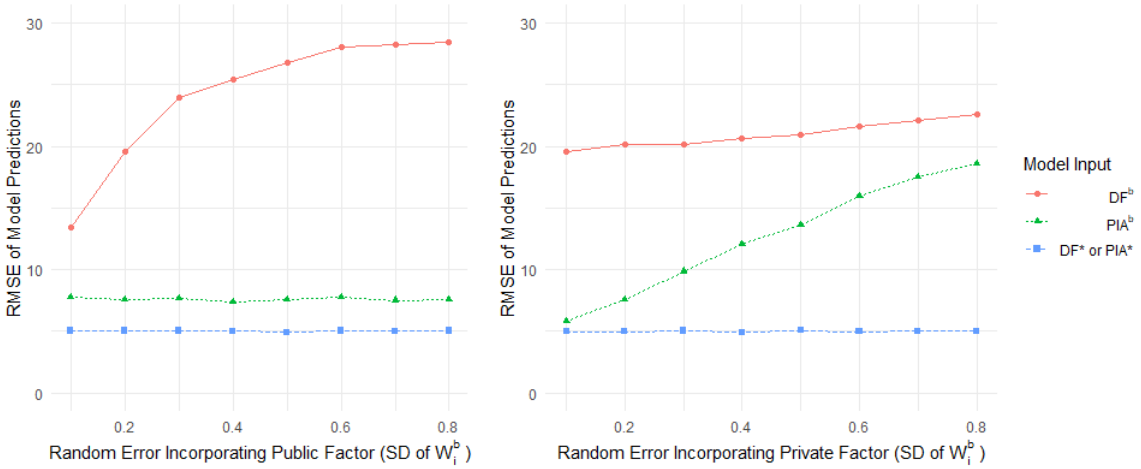
Random Error Incorporating Public Factors. To examine the effect of random error on the DF-PIA gap when incorporating public factors, we consider two cases that are identical except for the degree of variability in W_i^b for $i \in \mathcal{P}$. Specifically, we define $\widetilde{W}_i^b$ to be a mean preserving spread of W_i^b (Rothschild and Stiglitz 1970). The following result establishes that increasing the variability in how people incorporate public factors increases the DF-PIA gap:

PROPOSITION 3. *The DF-PIA gap is larger when $\widetilde{W}_i^b$ is used in (5), for $i \in \mathcal{P}$, instead of W_i^b.*

The idea behind Proposition 3 is as follows. Model-PIA^b remains the same when we add variability to $W_i^b, i \in \mathcal{P}$ because PIA responses are unaffected by the random error incorporating public factors. However, Model-DF^b is less accurate when $\widetilde{W}_i^b$ is used instead of $W_i^b, i \in \mathcal{P}$. DF responses are more variable when $\widetilde{W}_i^b$ is used, which makes it harder for the algorithm to learn the private factors. Combining these two observations implies that the DF-PIA gap increases.

Figure 1 shows the results from numerical simulations (see Appendix B for details). The left panel corresponds to Proposition 3. It varies the standard deviation of the public-factor random weight, holding constant the standard deviation of the private-factor random weight. Observe that the DF-PIA gap increases with the variability in the public-factor weight because the RMSE associated with Model-PIA^b remains constant, while the RMSE associated with Model-DF^b increases.

Figure 1 Numerical Simulation Illustrations of Propositions 3 and 4.



Random Error Incorporating Private Factors. We now turn to examining the effect of random error on the DF-PIA gap when incorporating private factors. We proceed as above, by considering two cases that are identical except for the degree of variability in W_i^b for $i \in \mathcal{I}$.

PROPOSITION 4. *The DF-PIA gap is smaller when* $\widetilde{W}_i^b$ *is used in (5), for* $i \in \mathcal{I}$*, instead of* W_i^b.

In contrast to Proposition 3, Proposition 4 shows that adding variability to how people incorporate private factors reduces the DF-PIA gap. Both Model-PIA^b and Model-DF^b lose accuracy as we add variability to $W_i^b, i \in \mathcal{I}$. However, the loss is more dramatic for Model-PIA^b. PIA's advantage of more directly eliciting the private information decreases as the random error incorporating private information increases.

The right panel of Figure 1 is the corresponding figure for Proposition 4. Observe that the DF-PIA gap decreases in the standard deviation of the private-factor random error term because while random error incorporating the private factor increases the RMSE under both Model-DF^b and Model-PIA^b, the increase is steeper in the latter.

Summary and Hypothesis. Proposition 3 shows that the DF-PIA gap increases in the random error incorporating public factors. Proposition 4 shows that the DF-PIA gap decreases in the random error incorporating private factors. Combined, they imply that the DF-PIA will be greater when adding random error incorporating public information than when adding the same amount of random error incorporating private information. Based on these results, we formulate our second hypothesis:

Hypothesis 2 *The location of random error moderates the DF-PIA gap. Specifically,*

(a) Inducing greater random error incorporating public information increases the DF-PIA gap.

(b) Inducing greater random error incorporating private information decreases the DF-PIA gap.

(c) Random error incorporating public information increases the DF-PIA gap more than random error incorporating private information.

3. Experiment 1: Elicitation via DF versus PIA

Experiment 1 is a simple direct test of Hypothesis 1.

3.1. Experimental Design

3.1.1. Task. Participants first reviewed 30 historical surgeries, each with information about the number of procedures, the anesthesia complexity score, and the resulting surgery duration. They then completed 50 rounds of surgery duration prediction. In each round, they were shown a new surgery's number of procedures and anesthesia complexity score. Then, they were asked a question about predicting its duration.

3.1.2. Conditions. Subjects were randomly assigned to one of two conditions: direct forecast ("DF") or private information adjustment ("PIA"). The only difference between these two conditions is that in each of the 50 rounds, DF participants answered the question "What is your forecast for the duration of this surgery? I think this surgery duration will be _____ minutes.", whereas PIA participants answered the question "The hospital system only has the first piece of information about this surgery—the number of procedures. You have additional information. To account for your additional information, how would you advise the hospital system to adjust its forecast for the duration of this surgery? I would advise the hospital system to increase/decrease (choose one) its forecasted surgery duration by _____ minutes." See Appendix, Figure E.1 for screenshots.

3.1.3. Simulating Surgery Duration. We used the following equation to simulate surgery duration: $Y_s = 60 + 20X_s^P + 10X_s^I + \epsilon_s$. Here, Y_s is the duration of surgery s; X_s^P denotes the number of procedures, an integer-valued public factor that has a uniform distribution between 1 and 10, inclusive; and X_s^I denotes the anesthesia complexity score, a private factor that has a uniform distribution between -5 and 5. Finally, ϵ_s follows a normal distribution with mean 0 and standard deviation 5. All participants observed the same 30 simulated historical surgeries. However, each participant observed a unique sequence of randomly generated surgeries for their 50 prediction rounds.

3.1.4. User Interface and Instructions. We programmed the user interface using the online software *SoPHIE* (Hendriks 2012). After receiving written instructions describing the task, participants were required to pass a three-question comprehension test before starting the experiment. They could review the instructions and retake the test until they answered all questions correctly. See Appendix, Figure E.2 for full instructions.

3.1.5. Pre-registration. For all experiments, we set our target sample sizes, exclusion criteria, and analysis plans a priori. We pre-registered to exclude participants who (1) did not complete the experiment or (2) put the same answer more than 90% of the time. We also pre-registered our dependent variable and analyses. We calibrate prediction algorithms using data from the participants' training set (first 35 rounds) and then use the algorithms to generate predictions on the test set (last 15 rounds). Our performance criterion is the RMSE of the predictions generated on the test set. The full pre-registration document for Experiment 1 is available at `https://aspredicted.org/blind.php?x=3e427n`.

3.2. Results

3.2.1. Participant and Response Summary Statistics. Undergraduate and graduate students at a large research university in the US were invited to participate through a behavioral laboratory subject pool recruitment system. Each participant received a $10 electronic gift card for completing the online study.

A total of 120 students participated. Following our exclusion criteria, we removed 8 participants who did not complete the experiment, leaving 112 for analysis (56 in each condition). Among the 112 participants, 75% were female, 88% were 18 to 24 years old, and 12% were 25 to 34 years old. The average of mean response to the question was 152.6 minutes (SD 30.8) in the DF condition and 12.1 minutes (SD 23.7) in the PIA condition.

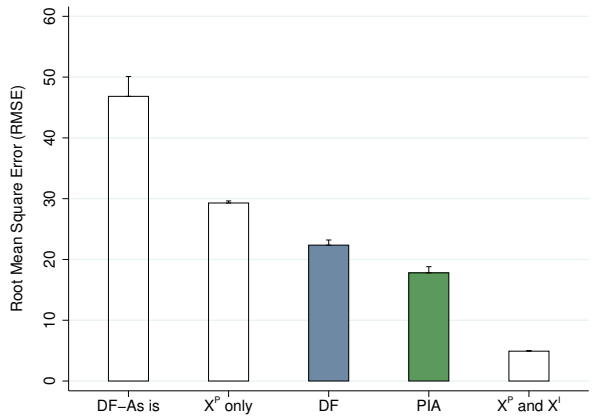
3.2.2. Algorithm Calibration and Prediction Accuracy Calculation. For each condition, we used the number of procedures, actual surgery duration, and participant response from all participants' first 35 rounds to calibrate prediction algorithms for surgery duration. The pre-registered linear regression model included participant dummies, number of procedures interacted with participant dummies, and participant response interacted with participant dummies. Table E.1 in the Appendix summarizes the prediction algorithms calibrated for each condition. We used the calibrated prediction algorithms to generate the predictions, $\hat{Y}_s$, for each surgery s in the last 15 rounds for each participant. We then computed $RMSE = \sqrt{\frac{1}{15}\sum_{s=1}^{15}(Y_s - \hat{Y}_s)^2}$ for each participant.

3.2.3. Testing Hypothesis 1. The average RMSE (averaged across all participants) was 22.4 (SD 6.3) in the DF condition and 17.8 (SD 7.6) in the PIA condition; see Figure 2(a). This difference of 4.6 is significant ($p = 0.0008$) and represents a 21% decrease. This result supports Hypothesis 1.

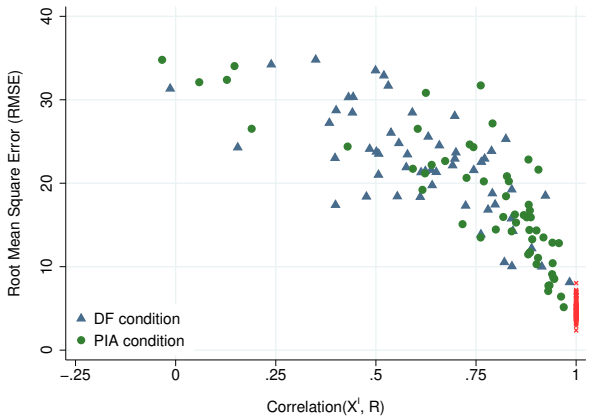
3.2.4. Other Benchmarks. Figure 2(a) also shows the performance of three other benchmarks:

"DF-As is" corresponds to using participant DF condition responses without any algorithms. Doing so results in an average RMSE of 46.8 (SD 24.5), significantly worse than when we use participant responses as inputs to algorithms.

Figure 2 Experiment 1: Performance Comparison.



(a) RMSE comparison. Means and standard errors are shown. X^P is the public factor, and X^I is the private factor.



(b) Correlation(X^I, R) versus RMSE. Each dot is one participant. R is defined as the residual of response after regressing it on X^P. Red x marks show the performance of the "X^P and X^I" model.

"X^P only" corresponds to using *only* the public information in the algorithm, without the use of any participant responses. Across the 112 participants, such an algorithm leads to an average RMSE of 29.3 (SD 3.5)—an improvement over "DF-As is" even though participants had access to the private information in addition to the public information. However, it is worse than the average RMSE of both the DF condition ($p < 0.0001$) and the PIA condition ($p < 0.0001$). In other words, participant responses added predictive value in the experiment.

Lastly, "X^P and X^I" corresponds to allowing prediction algorithms to directly observe the private information X^I and include it in prediction algorithms. In this experiment, it is equivalent to "consistent forecasters" and is a benchmark for the best performance possible. This algorithm results in an average RMSE across the 112 participants of 4.9 (SD 1.1).

3.2.5. Mechanistic Evidence. The theorized mechanism driving Hypothesis 1 is that PIA responses help the algorithm account for the private information better than the responses from the DF condition. To examine this mechanism, we calculate the correlation of the PIA responses with X^I and compare them with the correlation of the DF responses with X^I.

Naturally, because the PIA asks directly about the private information, the correlation between X^I and response was lower in the DF condition than in the PIA condition (0.39 versus 0.74). Next, we consider the correlation between X^I and R, where we define R as the residual of participant response after regressing it on X^P. That is, we take out the effect of public factor from each response to construct R. Note that if participants did not suffer from random error, then R would be perfectly correlated with X^I in both DF

and PIA conditions. In contrast, we find that it is less than 1 in both conditions. However, it is significantly higher in the PIA condition than in the DF condition (0.76 versus 0.62, $p = 0.0008$). In other words, PIA responses provide better information about X^I than DF responses.

Figure 2(b) illustrates the predictive accuracy versus the correlation value above for each participant. As expected, higher correlation between X^I and R leads to better prediction performance. There are more participants with high correlation in the PIA condition than in the DF condition, which contributes to the better performance of the PIA condition overall. The red "x" marks indicate the hypothetical perfectly-consistent benchmark, with no random error for each participant, which yields perfect correlation for both DF and PIA conditions. The deviation of the PIA and DF dots from the red marks illustrates the effect of human random error in participant responses.

3.3. MTurk Replication

We replicated the same experiment with MTurk workers. See `https://aspredicted.org/blind.php?x=yv2vs7` for the pre-registration document. While overall, the predictions from the experiment with MTurk workers were less accurate, the between-condition results replicated, providing evidence of robustness across different populations. Appendix C provides details on the replication as well as a comparison between the performances of university students and MTurk workers.

3.4. Discussion

Consistent with Hypothesis 1, Experiment 1 provides evidence that eliciting the PIA information instead of DF leads to better prediction algorithm performance. After the effect of public factor is taken out, participant responses are more correlated with the private factor. This tighter relationship helps prediction algorithms to incorporate private information, leading to better predictive performance. These results were replicated across university students and MTurk workers.

4. Experiment 2: Manipulating Information Complexity of Public versus Private Factors

Experiment 2 was designed to test Hypothesis 2, namely how the DF-PIA gap established in Experiment 1 is moderated by random error in incorporating public versus private factors. In addition, it provides a replication test of Hypothesis 1 using different surgery duration equations.

4.1. Experimental Design

The task was similar to that in Experiment 1. However, we changed the surgery duration equation, and we varied the number of factors by condition. We conjectured that greater information complexity induces greater random error. Therefore, we created higher complexity to induce more human random error by requiring that subjects aggregate multiple factors. Otherwise, to create lower complexity to induce less

random error, we automatically pre-aggregated multiple factors into a single representative factor for the participant.

Specifically, in the *Baseline* case, we pre-aggregated information so that there was only one public and one private factor, as in Experiment 1. However, we required that participants account for two public factors in the *Public Info Complex* case or two private factors in the *Private Info Complex* case. Thus, the experiment had a 2 (DF, PIA) by 3 (*Baseline*, *Public Info Complex*, *Private Info Complex*) between-subject experimental design.

The equations below show the underlying model we used for all conditions and the pre-aggregations we constructed to manipulate complexity by condition:

$$\begin{aligned} Y_s &= 150 + 10X_s^{P1} + 10X_s^{P2} + 10X_s^{I1} + 10X_s^{I2} + \epsilon_s \quad \text{(Underlying Model)} \\ &= 150 + 10X_s^{P1} + 10X_s^{P2} \quad + (50 + 20X_s^{I}) + \epsilon_s \quad \text{(Public Info Complex)} \\ &= 150 \quad + (50 + 20X_s^{P}) + 10X_s^{I1} + 10X_s^{I2} + \epsilon_s \quad \text{(Private Info Complex)} \\ &= 150 \quad + (50 + 20X_s^{P}) \quad + (50 + 20X_s^{I}) \quad + \epsilon_s \quad \text{(Baseline).} \end{aligned}$$

Here, X_s^{P1} and X_s^{P2} represent the two public factors. In the experimental task, they are the "procedure set-up requirements" and the "procedure complexity score," respectively. Symmetrically, X_s^{I1} and X_s^{I2} represent the two private factors. In the experimental task, they are the "anesthesia set-up requirements" and the "anesthesia complexity score," respectively. The random generation process for public and private factors was symmetric. For every surgery s, X_s^{P1} and X_s^{I1} were uniform random integers between 0 and 10, inclusive. X_s^{P2} and X_s^{I2} were uniform random numbers between -5 and 5 (rounded to the nearest tenth). We set $X_s^P = (X_s^{P1} - 5)/2 + X_s^{P2}/2$ and $X_s^I = (X_s^{I1} - 5)/2 + X_s^{I2}/2$, which establishes the above equalities. In the experimental task, they are a generic "procedure score" and "anesthesia score," respectively. See Appendix, Figure E.3 for screenshots. The pre-registration document for Experiment 2 is available at `https://aspredicted.org/blind.php?x=9uq8dw`.

4.2. Results

4.2.1. Participants, Participant Responses, and Prediction Algorithm. MTurk workers who were located in the US, had a Human Intelligence Task (HIT) approval rate of 99% or higher, and had 100 or more HITs approved were qualified to participate in the experiment. Participants who completed the experiment were paid $2 for participation. A total of 480 MTurk workers participated. Following the pre-registered exclusion criteria, we removed 174 individuals who did not complete the experiment and 54 participants who failed to correctly answer a four-question comprehension test on their first attempt. Among the 252 remaining participants, 42% were female, and 8% were 18 to 24 years old; 36%, 25 to 34; 31%, 35 to 44; 13%, 45 to 54; and 11%, 55 or over. Columns (1) and (2) of Table 1 provide the number of participants

and the average response in each condition. We developed prediction algorithms in the same way as in Experiment 1 (see §3.2.2). Table E.2 in the Appendix summarizes the prediction algorithms.

Table 1 Experiment 2: Summary of Experiment Results.

Information Type	Question Type	(1) N	(2) Response	(3) Corr(X^I, response)	(4) Corr(X^I, R)	(5) RMSE of test set	(6) DF-PIA gap
Baseline	DF	47	236.6 (30.7)	0.52 (0.21)	0.63 (0.25)	35.2 (13.9)	12.5***
	PIA	42	4.0 (19.1)	0.76 (0.23)	0.79 (0.22)	22.8 (11.4)	
Public Info	DF	41	241.9 (24.5)	0.28 (0.18)	0.42 (0.26)	41.2 (11.2)	19.9***
Complex	PIA	38	13.8 (33.9)	0.80 (0.27)	0.80 (0.27)	21.3 (14.0)	
Private Info	DF	47	237.7 (31.6)	0.66 (0.15)	0.70 (0.15)	30.6 (8.8)	1.7
Complex	PIA	37	48.8 (46.1)	0.72 (0.17)	0.73 (0.17)	28.9 (7.0)	

Note. Means and standard deviations (in parentheses) are shown. X^P is the public factor, and X^I is the private factor. In column (4), R is defined as the residual of response after regressing it on X^P. In column (6), DF-PIA gap is defined as the difference between the mean RMSEs of DF and PIA conditions. Column (6) also shows DF-PIA gap's statistical significance from a two-sample t-test for difference of means. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2.2. Robustness of Hypothesis 1. Columns (5) and (6) of Table 1 summarize the prediction performance in each of the six conditions. Consistent with Hypothesis 1, the average RMSE of all PIA participants was 32% lower than the average RMSE of all DF participants (35.4 versus 24.2, $p < 0.0001$). As shown in column (6) of Table 1, the DF-PIA gap was statistically significant at the 5% level in two of the three information conditions. In §3.2.5, we found that better performance is associated with higher correlation between participant response and X^I after the effect of X^P in the responses is taken out. Columns (3) and (4) of Table 1 provide the average correlation between X^I and response and the average correlation between X^I and R, defined as the residual of response after regressing it on X^P. As expected, the correlations are higher in the PIA conditions than in the DF conditions, which again provides mechanistic evidence for Hypothesis 1.

4.2.3. Testing Hypothesis 2. Hypothesis 2(a) predicts the DF-PIA gap to be greater under public-information-complex conditions than under baseline conditions. Consistent with this hypothesis, the gap was 19.9 under the public-information-complex conditions and 12.5 under the baseline conditions. This difference of 7.4 was statistically significant ($p = 0.036$; see Table 2).

Hypothesis 2(b) predicts the DF-PIA gap to be smaller under private-information-complex conditions than under baseline conditions. Consistent with this hypothesis, the gap was 1.7 under the private-information-complex conditions and 12.5 under the baseline conditions. This difference of 10.8 was statistically significant ($p = 0.002$; see Table 2).

Hypothesis 2(c) predicts the DF-PIA gap to be greater under public-information-complex conditions than under private-information-complex conditions. Consistent with this hypothesis, the gap was 19.9 under public-information-complex conditions and 1.7 under private-information-complex conditions. This difference of 18.2 was statistically significant ($p < 0.001$; see Table 2). These findings provide evidence that the benefit of PIA over DF is greater when public information is complex than when private information is complex.

One unpredicted pattern is that the RMSE in DF is smaller under the private-information-complex condition than under the baseline condition (30.6 versus 35.2, $p <= 0.056$). A plausible explanation is that, in addition to inducing more random error, splitting the private information into two factors causes participants to weight the private information more in general (see Fox and Clemen (2005)).

Table 2 Experiment 2: Performance Comparison.

	(1)
	Root Mean Squared Error
Information type (Base is Baseline conditions)	
Public Info Complex conditions	5.96* (2.43)
Private Info Complex conditions	-4.64* (2.35)
Question type (Base is DF conditions)	
PIA conditions	-12.49*** (2.42)
Interaction effects (Base is Baseline conditions × PIA conditions)	
Public Info Complex × PIA	-7.42* (3.52)
Private Info Complex × PIA	10.77** (3.48)
Constant	35.25*** (1.66)
N	252
R^2	0.27

Note. Column (1) is a linear regression model with RMSE as the dependent variable. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.3. Discussion

In addition to replicating Hypothesis 1 under different simulation parameters and information variables, Experiment 2 provided evidence that the location of random error matters in a manner consistent with Hypothesis 2. Specifically, eliciting human judgment via PIA instead of DF is helpful because of random error incorporating public information. Increasing random error incorporating private information reduces this benefit.

5. General Discussion

5.1. Summary

Our theoretical and experimental results suggest that in some situations, there is an opportunity to substantially improve the way prediction algorithms incorporate human judgment by applying a new PIA elicitation method, instead of the traditional method of DF. Under DF, humans contribute random error as they account for public information that the algorithm can already access. This random error hinders the algorithm's ability to infer the humans' private information. PIA avoids this hindrance by asking more directly about how much to adjust for the human's private information.

5.2. What Is Private Information and Why Does It Exist?

Pragmatically speaking, private information is any predictive information the human observes that the algorithm does not use. Because private information is context specific, rather than attempting to discuss exactly what it is, we find it constructive to discuss reasons why humans may have information that the algorithm does not use (i.e., why private information exists). We discuss these reasons in the context of our motivating example of predicting surgery durations: see Ibrahim and Kim (2019). At this hospital, the public information is the patient's electronic medical record as well as answers to specific questions from a standardized booking slip for surgery. All other predictive information the surgeon has is private.

5.2.1. Identification Challenges. Some theoretically easy-to-input information may be private simply because the system fails to request it. In our experiments, private information is easy to identify because the researcher knows all the data that exist in the environment and what data the human can access that the algorithm cannot. In practice, such an exercise is more difficult because if information is kept private from the algorithm, it may also be kept from the system designer. In other words, you cannot ask for what you do not know exists. For example, a surgeon may need special anesthesia equipment that requires additional set-up time, but there may be no place to indicate this information on the booking slip because the system designer was unaware of this special equipment.

5.2.2. Privacy Concerns and Integration Barriers. Even if the system designer knows that humans have certain private information that should be used in the algorithm, humans may be unwilling to input this information. For example, the surgeon may have a mental estimate of the likelihood of making a severe mistake during the procedure, but he/she may be unwilling to record that information for liability reasons. Similarly, certain information may be stored somewhere that is difficult to integrate. For example, a surgeon may know which technician is scheduled to support the surgery, but that information is stored in another, unintegrated database.

5.2.3. Codification Challenges. Despite advancements in "big data," it often remains impractical to input and store certain types or large quantities of predictive information into an organization's system. For example, one study reports that experienced doctors use nearly two million pieces of information to treat their patients (Pauker et al. 1976). While some of the two million pieces of information may be explicit knowledge—knowledge that can be easily articulated, codified, stored, and accessed—they are likely to be tacit knowledge, or knowledge that is difficult transfer, such as intuitive judgment; see Cowan et al. (2000) for a detailed discussion. As a result, inputting this information into the system will be costly and time consuming (Pollack et al. 2014), if not impossible.

5.2.4. Insufficient Training Data. Even if information has been inputted into the system, it may remain unused by the prediction algorithm due to a lack of sufficient historical data for training. Macario (2006) reports that 50% of surgeries have less than five previous cases with the same procedure and same surgeon during the preceding year. Zhou and Dexter (1998) report that only 32% of their surgeries had two or more previous cases with the same procedure and same surgeon. Ibrahim and Kim (2019) remove about 60% of the surgeries from their data collected over three years to keep only the surgeries that had 30 or more cases with the same procedure and same surgeon. The fact that each specialty, or even each procedure, has its own meaningful variables that are specific to the specialty or the procedure exacerbates the problem (Hosseini et al. 2015). Thus, the algorithm designer may intentionally choose to leave certain information as "private" because of lack of training data to make it useful.

5.3. What Should System Designers Do?

5.3.1. Consider Eliciting Human Input Even If Human Forecasts Are Inaccurate. Our study was motivated by a hospital administrator who, concerned with poor human forecasting performance, was considering using algorithms and fully eliminating surgeons from the surgery duration forecast process. Our results highlight the fact that even when human forecasts are significantly worse than algorithms head-to-head, system designers can potentially significantly boost the algorithms' performance by seeking human input. Thus, when implementing prediction algorithms, system designers should check to see whether human judgment can boost algorithm performance before fully replacing humans.

5.3.2. Try Adding a PIA-Type Question. When using human judgment as an algorithm input, we suggest adding a PIA-type question, especially when the public information is complex or the system designer knows that there exists simple private information. Note also that DF and PIA are not mutually exclusive. Thus, if the system already elicits the DF, one may choose to add the PIA and use both as inputs to the algorithm. Prompting people to think differently via DF and PIA may, in fact, help people better communicate private information to the algorithm (e.g., Herzog and Hertwig 2009), although doing so requires more effort.

5.3.3. Identify and Convert Private Information into Public Information. The results of this paper suggest that PIA mitigates the undesirable impact of human random error relative to DF, but not completely. Directly converting private information to public information will be superior to eliciting the PIA (e.g., see Figure 2) once enough training data have been collected. Thus, in addition to eliciting the PIA, we suggest attempting to learn what the underlying private information is behind humans' answers and altering the system to collect or directly elicit it moving forward. While it is unlikely that one will be able to fully eliminate private information in this manner, in some contexts, one may be able to eliminate enough private information to render the improvement due to PIA or DF negligible.

5.3.4. Experiment with and Revise the PIA Elicitation Format. Because the type of private information is context specific, the best way to write the PIA question is also likely to be context specific. Therefore, we suggest experimenting with and periodically revising how to write the PIA question. In Appendix D, we have made some limited progress on this issue via an experiment. We found that changing the format of the PIA question to make it easier for the human to translate from the domain of the private information to the domain of the question can enhance PIA's performance. For example, on the one hand, if the private information is a relative assessment of complexity, then structure the PIA question as an assessment relative to an average patient with the same public information. On the other hand, if the private information is a delay in minutes, then format the PIA question to be in minutes.

5.4. How Can Future Research Help Improve (or Disprove) This PIA Idea?

One limitation of our laboratory study approach is that it does not directly address the question of whether PIA will outperform DF in practice. Certainly, field experiments or even more realistic laboratory experiments can help address this question. Nevertheless, our initial studies suggest that more investigation into how best to write the PIA question may be beneficial before one can confidently implement it and assess its performance relative to DF. How should one decide whether to make the PIA question in a relative domain (e.g., relative to an average case) or an absolute domain (e.g., in dollars)? What is the best way to describe public information in specific practical contexts? Should one decompose the PIA into multiple parts based on known categories of private information? Does showing the algorithm's forecast before humans provide their PIA help or hurt? Are there algorithm aversion or incentive issues that need to be addressed before implementation?

We also recognize that there are several other open theoretical questions. We have assumed linear relationships and linear models in this paper. Intuitively, we believe the main directional predictions apply to other machine-learning algorithms and non-linear relationships. However, future work may verify whether our results do indeed generalize, which may lead to further insight. Furthermore, we have assumed a simple

model of random error that does not capture detail in how humans turn cues into predictions, where exactly the random error occurs, or how the format of the PIA question might matter. Another potentially fruitful direction is to incorporate further psychologically descriptive detail into a behavioral model of prediction.

In conclusion, field work, laboratory experiments, and behavioral models are all important for enhancing the understanding and use of PIA questions. It is our hope that by defining the PIA idea and documenting its potential improvement experimentally, we stimulate research that drives improvement in practice.

Acknowledgments

The authors thank Felipe Osorno, Joan Brown, Dr. Anthony W. Kim, and Dr. Sang W. Lee for helpful discussions and advice. The authors are grateful to the editors and reviewers for their constructive and insightful comments, which helped improve the paper significantly. The authors also thank the seminar and conference participants at University of North Carolina at Chapel Hill OM Healthcare Research Workshop, 2019 INFORMS Healthcare Conference, 2019 INFORMS Annual Meeting, and Wisconsin School of Business OIM Seminar.

References

Alexander Jr JC (1995) Refining the degree of earnings surprise: A comparison of statistical and analysts' forecasts. *Financial Review* 30(3):469–506.

Arvan M, Fahimnia B, Reisi M, Siemsen E (2019) Integrating human judgement into quantitative forecasting methods: A review. *Omega* 86:237–252.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model+ 50% manager. *Management Science* 36(8):887–899.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management* 13(1):2–19.

Carbone R, Andersen A, Corriveau Y, Corson PP (1983) Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science* 29(5):559–566.

Childers CP, Maggard-Gibbons M (2018) Understanding costs of care in the operating room. *JAMA Surgery* 153(4):e176233–e176233.

Cowan R, David PA, Foray D (2000) The explicit economics of knowledge codification and tacitness. *Industrial and Corporate Change* 9(2):211–253.

Davis AM, Katok E, Kwasnica AM (2014) Should sellers prefer auctions? A laboratory comparison of auctions and sequential mechanisms. *Management Science* 60(4):990–1008.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International journal of forecasting* 25(1):3–23.

Fox CR, Clemen RT (2005) Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science* 51(9):1417–1432.

Goodwin P (2000) Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting* 16(2):261–275.

Hendriks A (2012) SoPHIE - Software platform for human interaction experiments. *Working paper, University of Osnabrueck* .

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science* 20(2):231–237.

Ho TH, Zhang J (2008) Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science* 54(4):686–700.

Hosseini N, Sir MY, Jankowski C, Pasupathy KS (2015) Surgical duration estimation via data mining and predictive modeling: a case study. *AMIA Annual Symposium Proceedings*, volume 2015, 640 (American Medical Informatics Association).

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Ibrahim R, Kim SH (2019) Is expert input valuable? The case of predicting surgery duration. *Seoul Journal of Business* 25(2).

Kahneman D, Rosenfield A, Gandhi L, Blaser T (2016) Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* 36–43.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Available at SSRN 3619085* .

Kim SH, Chan CW, Olivares M, Escobar G (2015) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* 61(1):19–38.

Kremer M, Siemsen E, Thomas DJ (2016) The sum and its parts: Judgmental hierarchical forecasting. *Management Science* 62(9):2745–2764.

Lawrence M, Goodwin P, O'Connor M, Önkal D (2006) Judgmental forecasting: A review of progress over the last 25 years. *International Journal of forecasting* 22(3):493–518.

Lawrence M, O'Connor M, Edmundson B (2000) A field study of sales forecasting accuracy and processes. *European Journal of Operational Research* 122(1):151–160.

Macario A (2006) Are your hospital operating rooms "efficient"? A scoring system with eight performance indicators. *Anesthesiology: The Journal of the American Society of Anesthesiologists* 105(2):237–240.

Mannes AE, Larrick RP, Soll JB (2012) The social psychology of the wisdom of crowds. Krueger JI, ed., *Frontiers of social psychology. Social judgment and decision making*, 227–242 (Psychology Press).

Osadchiy N, Gaur V, Seshadri S (2013) Sales forecasting with financial indicators and experts' input. *Production and Operations Management* 22(5):1056–1076.

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. *Management Science* 65(5):2291–2309.

Pauker SG, Gorry GA, Kassirer JP, Schwartz WB (1976) Towards the simulation of clinical cognition: taking a present illness by computer. *The American Journal of Medicine* 60(7):981–996.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* 61(8):1741–1759.

Pollack AH, Tweedy CG, Blondon K, Pratt W (2014) Knowledge crystallization and clinical priorities: evaluating how physicians collect and synthesize patient-related data. *AMIA Annual Symposium Proceedings*, volume 2014, 1874 (American Medical Informatics Association).

Rothschild M, Stiglitz JE (1970) Increasing risk: I. A definition. *Journal of Economic Theory* 2(3):225–243.

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management* 10(4):566–589.

Surowiecki J (2005) *The wisdom of crowds* (Anchor).

Tong J, Feiler D (2017) A behavioral model of forecasting: Naive statistics on mental samples. *Management Science* 63(11):3609–3627.

Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science* 19(7):645–647.

Zhou J, Dexter F (1998) Method to assist in the scheduling of add-on surgical cases-upper prediction bounds for surgical case durations based on the log-normal distribution. *Anesthesiology: The Journal of the American Society of Anesthesiologists* 89(5):1228–1232.